International Journal of Information and Communication Sciences
2018; 3(3): 88-95
http://www.sciencepublishinggroup.com/j/ijics
doi: 10.11648/j.ijics.20180303.12
ISSN: 2575-1700 (Print); ISSN: 2575-1719 (Online)

Text Clustering Incremental Algorithm in Sensitive Topic Detection

Yuejin Zhang[1], Jiajia Zhang[2], Dongmei Zhao[3, *]

[1]Cyber Security Department, Municipal Cyberspace Administration, Beijing, China

[2]Eliot K-8 Innovation School, Boston, America

[3]Department of Electronic Commerce, China Agricultural University, Beijing, China

Email address:

zhaodongm@vip.163.com (Dongmei Zhao)

[*]Corresponding author

To cite this article:

Yuejin Zhang, Jiajia Zhang, Dongmei Zhao. Text Clustering Incremental Algorithm in Sensitive Topic Detection. *International Journal of Information and Communication Sciences*. Vol. 3, No. 3, 2018, pp. 88-95. doi: 10.11648/j.ijics.20180303.12

Received: August 28, 2018; **Accepted**: September 27, 2018; **Published**: October 30, 2018

Abstract: With the rapid development of Internet technology, the influence of online consensus continues to expand. How to quickly and effectively discover sensitive topics and keep track of those topics has become an important research recently. Text clustering can aggregate news texts with the same or similar content to achieve the purpose of discovering topics automatically. Make improvement to clustering algorithm according to different media types is the main research direction. Although the existing typical clustering algorithms have certain advantages, they all face constraints on data size and data characteristics in specific applications. There is no existing algorithm can fully adapt to these characteristics. Although the application of more Single-pass algorithms in the (TDT) field can realize the discovery and tracking of topics, there are disadvantages of poor accuracy and slow speed under massive data. According to the dynamic evolution characteristics of online consensus, this paper proposes an incremental text clustering algorithm based on Single-pass, which optimizes the clustering accuracy and efficiency of massive news. Based on the real online news texts from the online consensus analysis system, we conduct an experiment to test and verify the feasibility and effectiveness of the algorithm we proposed. The result shows that the new algorithm is much more efficient compared to the original Single-pass clustering algorithm. In the real application, the new incremental text clustering algorithm basically meets the real-time demand of online topic detection and has a certain practical value.

Keywords: Topic Detection, Online Consensus, Simhash Algorithm, Text Clustering, Incremental Algorithm, Single-Pass Algorithm

1. Foreword

Targeted to monitor and analyze online consensus is a research focus on current Natural Language Processing. Through crawling real-time web content from websites produce consensus frequently, the analysis system of online consensus offers incident warning, tracing analysis, consensus report to customers by discriminating content, topic detection, evaluating hot rank and analyzing sensitivity the web content crawled. Topic detection is the key process among the above processes which mainly depends on text clustering.

Text clustering [1-3] aggregates same or similar content of news texts to achieve detecting topics automatically. In this paper, we discuss and research an incremental text clustering algorithm for sensitive topic detection in an online consensus analysis system. Considering the text clustering process for topic detection must give consideration both on clustering result and efficiency based on actual requirements, in this paper we propose an incremental text clustering algorithm which based on Simhash.

2. Related work

Topic detection derives from Topic Detection and Tracking, TDT [4, 5]. So far, the algorithm applies to topic detection mainly based on incremental clustering algorithm. Single-pass algorithm which based on near-neighbors, is a classic algorithm [6]. The algorithm has simple principle and rapid

calculation without setting class number in advance. However, the algorithm has low accuracy of clustering result and relies on the sequence of data input. The clustering speed will decrease along with the growth of total texts of dataset. Among the improved algorithms, Yin Fengjing [7] proposed ICIT algorithm and Lei Zhen [8] proposed improved incremental algorithm. As the incremental text clustering algorithm can utilize the last clustering result, avoiding re-clustering the whole text cluster, improves the clustering efficiency greatly, has the possibility to satisfy the real-time performance in the requirements of topic detection process.

The dimension reduction process on feature vector avails to decrease the problem scale to improve the classification efficiency. Existing research mainly relies on the dimension reduction of Hash function [9, 10]. Text similarity calculation [11]. which depends on similar Hash fingerprint, can get a similarity theoretically approximate to the result of calculating feature vector directly, consequently approximate get a similarity of source text. The preprocessing decreases the scale and complexity of similarity calculation and is easy to apply to large-scale text clustering analysis in actual application scenario. Therefore, in this paper we will research on news text clustering analysis based on Simhash fingerprint.

The particularity of the application scenario of topic detection brings the following problems and difficulties:

1. The indeterminacy of topic numbers of news data causes the classes numbers face indeterminacy accordingly. Therefore, traditional clustering algorithm which need to assign the classes numbers, can not apply to the scenario immediately.
2. Topic detection applies to mass news data. The timeliness ensures the utility value of topic detection. Thus, it requires the algorithm must simple and efficient. Some clustering algorithms with high-quality, but complexity in calculation or space could not be able to apply.
3. The news data in topic detection increase all the time. Re-clustering the whole data when text data increase although ensure the clustering accuracy but fails to real-time clustering requirement. Introduce incremental clustering algorithms can utilize the last clustering result but will not influence the former clustering result when data incremental increase. The order of the text into the preprocessing will influence the clustering result.

The problems and difficulties above refer to the process of topic detection faces the increasing mass news data and the objective requirement of real-time processing. Thus, in this paper, the research priority is how to improve the efficient of text clustering process as far as possible.

3. Incremental Text Clustering Based on Simhash

3.1. The Whole Scheme Design

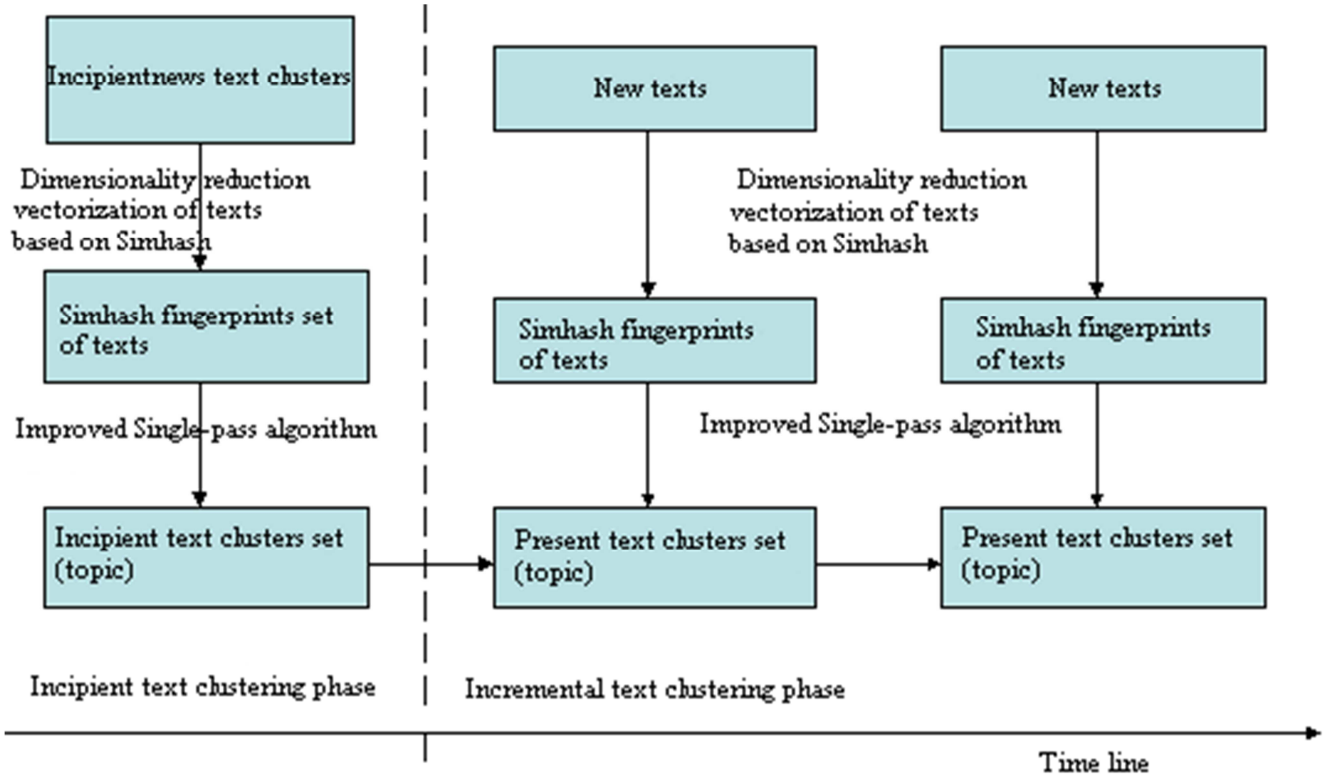

Figure 1. *Incremental text clustering based on Simhash.*

The whole executive process of incremental text clustering based on Simhash in this paper is shown as Figure 1. There are three phases in data processing for single news text which participates in clustering: Text vectorization, dimension reduction process for high-dimensional text vector, clustering analysis for text vector after dimensionality reduction and get clustering results.

3.2. Text Vectorization Based on News Text Feature

Based on Bag-of-Word model, in this paper, firstly, we propose to use ICTCLAS for word segmentation. Secondly, for character term set from word segmentation, we remove Stop Words, then keep the notional words. Thirdly, calculate TF-IDF value of every single character term to get characteristic vector of texts (IDF of word based on value tablet of character term calculated in advance). Finally, we adjust the weight of character term based on news text feature above. Considering the feasibility in the actual application scenario, firstly, weighting character term of the first-void must on the condition that we extract text from web content accurately. Secondly, we must label fine sorted word class and category through word segmentation to weighting time, place and name. Considering the two processes above cannot be ensured in actual application, in this paper we do not discuss relevant character term weighting but get a title vector by calculate word segmentation of title, then weighting the title vector.

3.3. Dimensionality Reduction Based on Simhash

In this paper we process dimensionality reduction on text vector based on Simhash. Simhash can calculate vectorization text to create corresponding fixed length Hash, called text fingerprint. Simhash ensures that when similar or same text input, we can get the similar or same fingerprint. Thus, by calculating the similarity of the fingerprints, the similarity of the source text can be measured and reflected. As follows is the process of calculating the Simhash fingerprint of the TF-IDF.

Step 1: There are T digits in Simhash, initialize every digit as 0;

Step 2: Calculate corresponding T digits Hash code of each characteristic term of text vector through traditional Hash function.

Step 3: For each digit of the T digits Hash code of characteristic term, if the digit is 1, then weighting the present characteristic term in corresponding place in the Simhash. If the digit is not 1, then subtract corresponding weight.

Step 4: For every digit of the T digits Simhash we get at last, if the digit is greater than 0, then set the digit to 1. Otherwise, set the digit to 0. Now we get the T digits two-value text fingerprint.

Step 5: We use Hamming distance to measure the similarity between the two text Simhash fingerprints through Simhash algorithm.

3.4. Improved Single-Pass Algorithm Based on News Timeliness

On the premise that news texts are input by release time, the clustering result of Single-pass algorithm can meet our requirement. Thus, in this paper we propose improvement methods towards Single-pass algorithm aiming at low accuracy of clustering result and clustering speed decreases within the increase of text set

(1) Calculating similarity between text and text cluster by average link.

Because of the easing strategy of single link adapted in calculating the similarity between text and text cluster, there are many large texts clusters content lots of texts in the clustering result, which influences the accuracy of clustering result. In this paper, we propose to calculate the similarity between text and text cluster by average link, which means to calculate the average value of the similarity between texts and texts of a text cluster. When the average value reaches a default threshold value, the text will be put into the cluster. In experimental evidence, the application of average link can decrease large text cluster appearing in clustering result and the accuracy rate also rises.

(2) Limit the numbers of text clusters when compared the similarity with new text incremental text clustering

The primary cause that clustering speed of Single-pass algorithm decreases when the text set increases is: according to the process of the algorithm, to calculate the similarity, all the texts of the text cluster must be compared with the new text one by one. The time complexity of the incremental text clustering process is O(n). Considering the news texts of the same news topic concentrate in a sustained period generally, then for a new news text (including recent topic), the probability of the cluster singly or the probability that the new news text is in the recent news (topic) outclasses the probability that the new news text is not in the recent news (topic).

Therefore, we introduce the parameter of Time Window. We define a news text cluster in the Time Window as an active text cluster. The new text will be in incremental text clustering with present active texts to avoid calculating the similarity with each text in the text set one by one, which solve the problem that the clustering speed of Single-pass algorithm decreases when the text set increases.

Based on the two improvement, the incremental text clustering process of the Single-pass algorithm is shown as Figure 2:

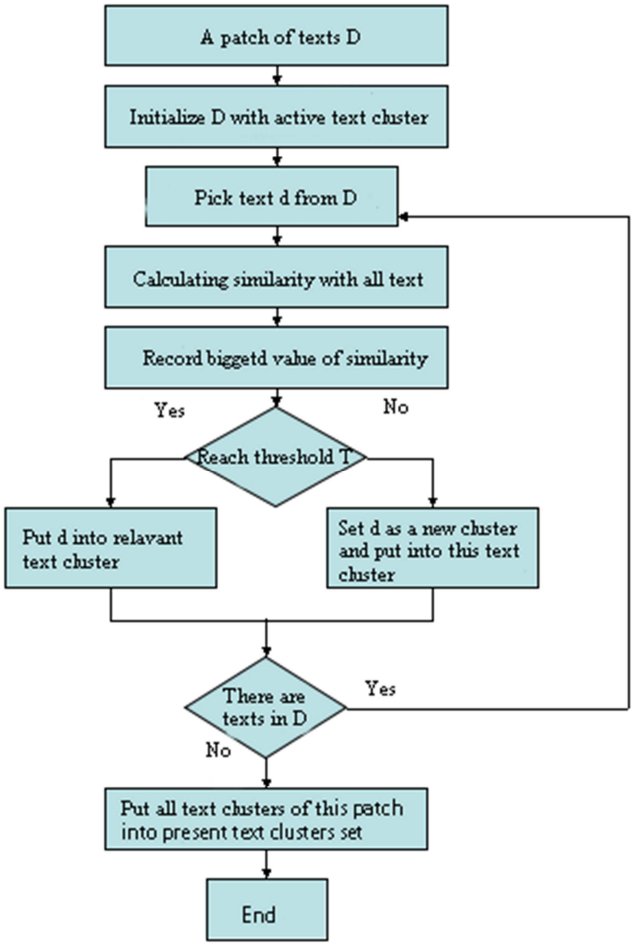

Figure 2. *The incremental clustering process of improved Single-pass algorithm.*

The improved Single-pass algorithm not only sets a threshold value parameter T, but also adds time window parameter K (unit: day) to limit the numbers of text clusters when compared the similarity with new text incremental text clustering. According to experimental result and the effect in actual application, we set the value of K as 3, which means before the incremental text clustering of each batch of texts, pick active text clusters in three days (including news texts published in latest 3 days) to initialize the text clusters set.

Because a sustaining concerned news topic produces relevant reports, which keep the text cluster active, when the time window parameter K is correspond to reality and reasonable, if the new text belongs to a historical text cluster, the reason is either the influence of the potential topic is weak, or the topic it presents is concerned again. According to the text algorithm, we will set a cluster for the text separately. In addition, for the initial clustering phase in the process of clustering, the Single-pass algorithm in this paper can also be applicated in the process of incremental text clustering。

3.5. Time Complexity Analysis

In this improved Single-pass algorithm, suppose the amount of each batch of texts is m, and the text of current batch only compared to the text cluster of current batch of text clusters for similarity, while the current text cluster only includes the active text cluster from former k days to current time, (suppose the amount of these text clusters is c, the text maximal amount under each text cluster is z), then conduct incremental clustering to single text, the time complexity will be no more than O(m+c*z), of which m is preset parameter, value of c*z is irrelevant to text total amount n, and it will be far less than n when text total amount n grows. So we consider O(m+c*z)≈O(1) approximately, it shows the incremental clustering process speed of the improved Single-pass algorithm will maintain constant, solving the problem that the speed of the original Single-pass algorithm will slow down when value n increases basically.

4. Experiment and Assessment

4.1. Statistic

Experiment statistics include all the crawled internet text related to “anti-embezzle, anti-corruption”, altogether 25272 pieces, as seen in Table 1.

Table 1. *Text data source statistics.*

Source	Quantity
Web news	11338
Forum	5564
Microblog	405
Blog	1805
WeChat	5930
Other	230
Total	25272

We manual picked out more than 30 topics which have more related texts, including 8901 pieces text.

4.2. Assessment of Clustering Algorithms

This experiment is comparing the clustering results of using different algorithm to cluster, using 64-bit Simhash fingerprint, similarity threshold value parameter T as 8.

4.2.1. Evaluation Criterion of Clustering Result

To evaluate these text clustering results objectively, generally it requires a previous manual classification of the texts in experimental corpus database and takes the manual classification result as comparison sample with clustering result. Because the text statistics in the experiment is too massive to classify each text, the experiment only evaluates the clustering results of the texts belonging to manual classification text clusters, including 30 topics, 8901 texts. We use Value F-Measure to evaluate the quality of the clustering results.

4.2.2. The Impact of Different Link Strategy on Clustering Result

Part of clustering results statistic is in table 2 by using Single-pass algorithm. n_e represents anticipated text amount under topic e with manual classification, n_r represents actual text amount under topic e during clustering, h (e, r) represents the anticipated text amount that clustering results match with.

Table 2. *Clustering result of original Single-pass algorithm(part).*

Topic e	n_e	n_r	h(e, r)	precision(e, r)	recall(e, r)	F_1(e, r)
Selling official position	911	950	886	93.26%	97.25%	95.21%
Corrupt officials hid money	134	504	126	25.00%	94.03%	39.50%
Investigate provincial official	395	451	383	84.92%	96.96%	90.54%
Case Liu and Gu	385	543	380	69.98%	98.70%	81.90%
Circle of official & merchant	780	977	776	79.43%	99.49%	88.33%
Sound section chief out	439	594	436	73.40%	99.32%	84.41%

Using Single-pass algorithm based on average link, part of the clustering results statistic is in table 3.

Table 3. *Clustering result of improved Single-pass algorithm based on average link(part).*

Topic e	n_e	n_r	h(e, r)	precision(e, r)	recall(e, r)	F_1(e, r)
Selling official position	911	869	842	96.89%	92.43%	94.61%
Corrupt officials hid money	134	126	120	95.24%	89.55%	92.31%
Investigate provincial official	395	224	203	90.63%	51.39%	65.59%
Case Liu and Gu	385	375	368	98.13%	95.58%	96.84%
Circle of official & merchant	780	605	582	96.20%	74.62%	84.05%
Sound section chief out	439	388	381	98.20%	86.79%	92.14%

The F value of Single-pass algorithm clustering results based on average link strategy and the original Single-pass algorithm is in Figure 3.

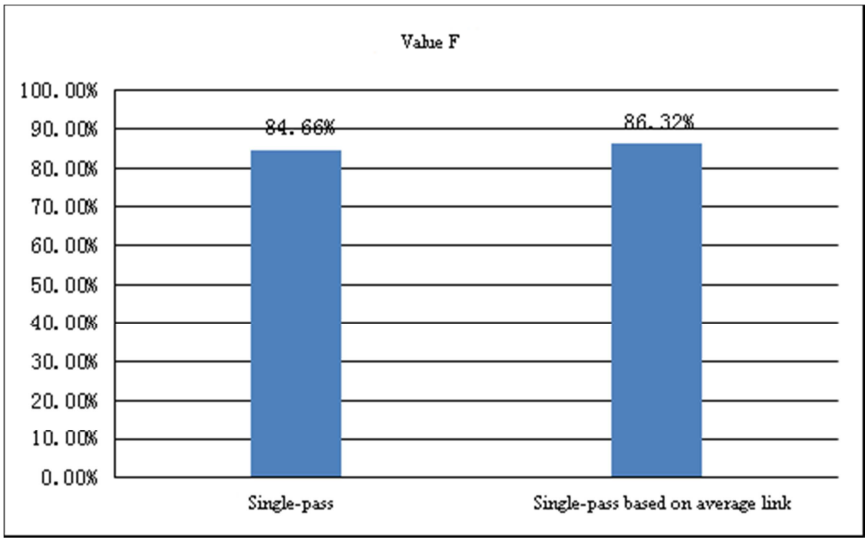

Figure 3. *F values of different Single-pass clustering results based on different link strategies.*

Figure 3 shows that the clustering results in comparison of Single-pass clustering algorithm based on average link strategy and the original Single-pass clustering algorithm, the F value Single-pass clustering algorithm based on average link strategy has little promotion by 1.96%. But in comparison of the accuracy and recall rate of the clustering results of the topics in Table 2 and Table 3, it's clear shows that the accuracy rate of Single-pass algorithm based on average link strategy has much more improvement than the original Single-pass algorithm. The original Single-pass algorithm has higher recall rate but lower accuracy (no more than 85%). It reflects that the original Single-pass algorithm has poor clustering accuracy. On the contrary of that, the Single-pass algorithm based on average link has higher clustering accuracy but lower recall rate. The recall rete some topic's is only around 50%.

4.2.3. Time-Consuming Comparison of the Clustering

First, taking value 0.25, 1, 3, 7 of time window K severally, then calculate the execution time of the clustering process (the execution time of text vectorization and dimensionality reduction is excluded). Second, compared the execution time with the execution time of the original Single-pass, as well as the execution time of the Single-pass based on average link without the time window K, taking average execution time of 6 operations of clustering for the above condition, the statistics is shown in Figure 4.

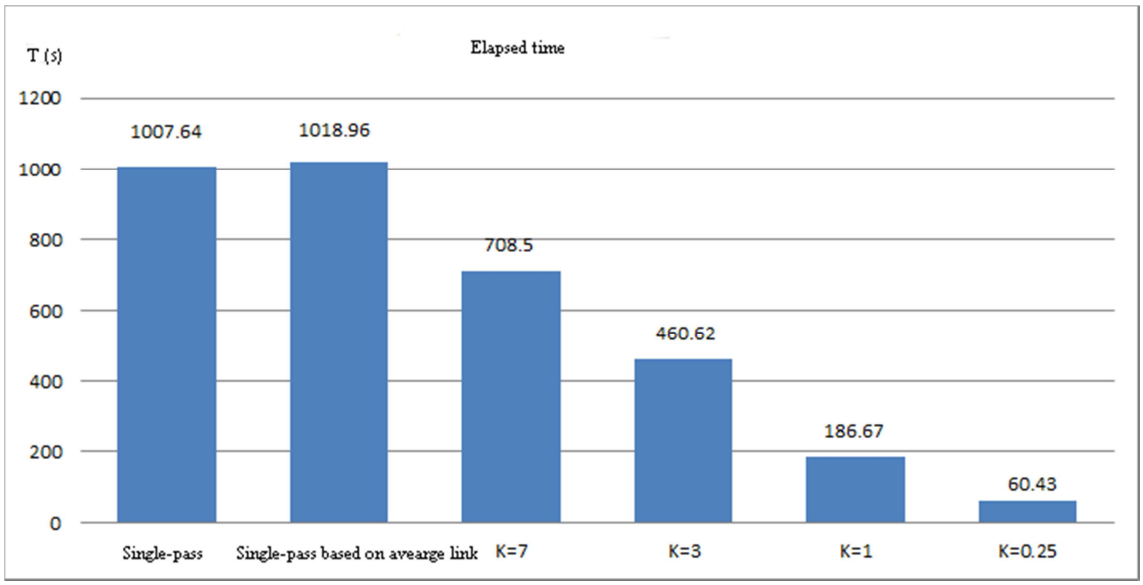

Figure 4. *Clustering process time when using different K values.*

The Figure 4 shows the execution time of Single-pass algorithm based on average link is almost the same as the execution time of Single-pass algorithm based on based on minimal link strategy in general. The more execution time of Single-pass algorithm based on average link mainly attributed to text need to be calculated with all texts in the text clusters. In comparison of the execution time of clustering in different value of K, it proves that as once the value K decreases, the execution time of clustering shortens observably. The reason is in the improved Single-pass algorithm which based on text, as time window K decreases, the active limited text clusters which match the value K in current clusters also decreases, and the texts of the similarity calculation with new texts in incremental clustering process also decreases.

4.2.4. Experiment Results Analysis

Experiment result indicates that the news texts of most topics releases in a concentrated time obviously, declaring that the improvement of introducing time window parameter based on news timeliness is reasonable and doable. In comparison of the clustering results of Single-pass based on average link strategy and the original Single-pass, although the value F is close, the Single-pass based on average link strategy has higher accuracy obviously. Analyzing the execution time by the different value K in time window parameter, due to the reduced text similarity calculation by introducing the time window parameter K in the process of incremental text clustering, it takes less time than original Single-pass algorithm. Furthermore, in comparison of the clustering results of different time window parameter K, when K is great than or equal to 3, its result is almost same as in only based on average link strategy Single-pass, namely, it has basically no influence on its clustering effect. Based on above analysis, we can conclude that this article suggested the improved Single-pass algorithm based on news timeliness has higher clustering efficacy than the original algorithm.

4.3. Realistic Application

About the topic detection process in online consensus analysis system, we can applicate the incremental text clustering algorithm based on Simhash, using 64-bit text Simhash fingerprint, setting Similarity threshold parameter T to 8, Time window parameter K to 3. Figure 8 demonstrates the interface of customized hot topics page in online consensus analysis system. User can customize his hot topic/news page by limiting the search criteria based on own focuses, such as news resource, type, topic keyword etc. Figure 5 demonstrates two topic news pages “Hubei disclose” and “Hubei corruption” created by test account.

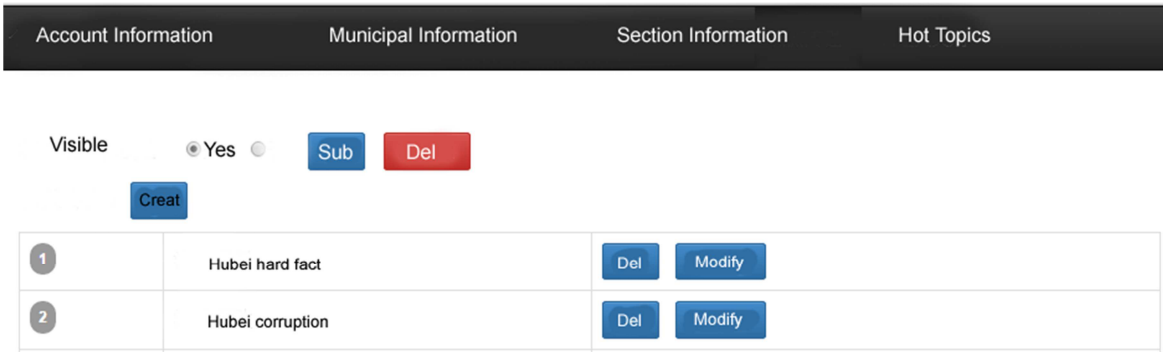

Figure 5. *The web page of hot topic customization.*

Consensus Report

Topics List

1	A prosecutor in Hubei province has been accused of cheating (538)	SINA	20:23,07-05-2015
2	A hospital in Wuhan has been destroyed forcedly (460)	Zuojiang News	16:54,07-05-2015
3	There is a wide gap between the fall of corrupt officials (221)	MOP	17:11,07-05-2015
4	134 violations of corruption regulations (171)	Wechat	17:23,07-05-2015
5	Public drinking water in metro stations is contaminated (142)	Baidu	11:56,07-05-2015
6	That's the truth about corrupt officials (94)	Zuojiang News	20:39,07-05-2015
7	A man washes his feet while drinking water on a subway in Wuhan (92)	Xici	20:45,07-05-2015
8	The ministry of culture cut the prize by 60 percent (82)	Wechat	19:35,07-05-2015
9	One hundred thousand catties of fish fry died from disease in... (67)	Dayang News	15:33,07-05-2015
10	Central Commission for Discipline Inspection of the CPC... (59)	Wechat	18:04,07-05-2015

« 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 »

Figure 6. *An example of the real time clustered topics.*

User can choose certain topic to check the related news list. This list comes from the cluster resourcing from text clustering. Figure 7 demonstrates the latest 10 news of the topic "a hospital in Wuhan got eviction".

News List

1	A hospital in Wuhan was forced to demolish	Zuojiang Daily	16:54,07-05-2015
2	Tens of millions invested private hospital in Wuhan have been…	Nanyang News	08:24,07-05-2015
3	Wuhan invested tens of millions of hospital by forced demolition of …	Zhonghua News	21:15,06-05-2015
4	The hospital was taken down by aliens	Xinggangfazhan	10:27,06-05-2015
5	Pretending not to know after the hospital was demolished forcedly	Nanyang News	08:20,06-05-2015
6	Wuhan private hospital forcedly demolished in middle night...	SCMP	04:47,06-05-2015
7	Tens of millions invested hospital in Wuhan has been demolished...	SINA	19:05,05-05-2015
8	Tens of millions invested hospital in Wuhan has been forcedly...	Apollo	19:04,05-05-2015
9	Private hospital in Wuhan has been forcedly demolished...	Zhonghua News	12:18,05-05-2015
10	The hospital was demolished and the decision makers could...	Yunnan News	09:01,05-05-2015

« 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 »

Figure 7. *An example of the news list in a clustered topic.*

In perspective of practical application effect, this Article proposed an incremental text clustering algorithm which based on Simhash basically achieves the anticipated demand of real-time topic detection in the system, proving high application value.

5. Conclusion

Focusing on the application scenario of topic detection in an online consensus analysis system, this Article proposes a new incremental text clustering algorithm which based on Simhash, to fulfil the realistic requirement that topic detection can be both accurate and speedy in text clustering algorithm. The experiment result verified the practicability and the effectiveness of this algorithm, which also proves its application value in realistic consensus analysis system.

References

[1] Chen Ning. Research in clustering algorithm of data excavation [D]. Mathematics and systematic science in CAS, 2001.

[2] Chen C C, Chen Y T, Sun Y, et al. Life cycle modeling of news events using aging theory[M]//Machine Learning: ECML 2003. Springer Berlin Heidelberg, 2003: 47-59.

[3] Liu Yuanchao, Wang Xiaolong, Xu Zhiming etc. Text clustering Summary [J]. Chinese Information Journal, 2006, 20(3): 55-62.

[4] J Azzopardi, C Staff. Incremental Clustering of News Reports. Algorithms, 2012, 5 (3): 364-378.

[5] Company, Suizhou 441300, China. Applied research of text clustering algorithm in network monitoring public opinion [J]. Electronic Design Engineering, 2013-01.

[6] Yang Y, Carbonell J, Brown R, et al. Learning approaches for detecting and tracking news events [J]. Intelligent Systems & Their Applications IEEE, 1999, 14(4): 32-43.

[7] Yin Fengjing, Xiao Weidong, Gebing etc. An incremental text clustering algorithm facing to internet topic detection [J]. Computer Application Research, 2011, 28(1): 54-57.

[8] Lei Zhen, Wu Lingda, Lei Lei etc. The incremental parameter K in average value method of initial class center and its application in news exploration [J]. Intelligence Academic Journal, 2006, 25(3): 289-295.

[9] X Yi, X Zhao, N Ke, F Zhao etc. An improved Single-Pass clustering algorithm internet-oriented network topic detection. International Conference on Intelligent Control & information processing, 2013: 560-564.

[10] M Mittal, RK Sharma, VP Singh. Modified single pass clustering with variable threshold approach. «International Journal of Innovative Computing information & control Ijicic», 2015, 11 (1): 375-386.

[11] Charikar M S. Similarity estimation techniques from rounding algorithms [C]//Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. ACM, 2002: 380-388.